W70–W74 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl043

FOLD-RATE: prediction of protein folding rates from amino acid sequence

M. Michael Gromiha*, A. Mary Thangakani[1] and S. Selvaraj[2]

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan, [1]Advanced Technology Institute Inc., Tokyo, Japan and [2]Department of Bioinformatics, Bharathidasan University, Tiruchirappalli 620 024 TN, India

Received December 21, 2005; Revised and Accepted February 3, 2006

ABSTRACT

We have developed a web server, FOLD-RATE, for predicting the folding rates of proteins from their amino acid sequences. The relationship between amino acid properties and protein folding rates has been systematically analyzed and a statistical method based on linear regression technique has been proposed for predicting the folding rate of proteins. We found that the classification of proteins into different structural classes shows an excellent correlation between amino acid properties and folding rates of two and three-state proteins. Consequently, different regression equations have been developed for proteins belonging to all-α, all-β and mixed class. We observed an excellent agreement between predicted and experimentally observed folding rates of proteins; the correlation coefficients are, 0.99, 0.97 and 0.90, respectively, for all-α, all-β and mixed class proteins. The prediction server is freely available at http://psfs.cbrc.jp/fold-rate/.

INTRODUCTION

Prediction of protein folding rates from amino acid sequences is one of the most important challenges in computational and molecular biology (1). Several investigations have been carried out to understand/predict the folding rates of proteins from protein 3D structures. These studies include the concept of contact order (2), first principles of protein folding (3), long-range order (4), elementary statistical model (5), combination of contact order and stability (6), number of native contacts (7), total contact distance (8), topomer search model (9), the topological properties of protein conformation (10), neural networks based on contact order, long-range order and total contact distance (11), amino acid properties (12), chain length (13), size (14), helix parameter (15) and native state geometry (16). Recently, different methods have been proposed for predicting protein folding rates from amino acid sequence, secondary structure and structural class information (17–20).

The folding of a protein is mainly dictated by inter-residue interactions, which are influenced by physical, chemical, energetic and conformational properties of amino acid residues (21). Further, amino acid properties have been successfully used for understanding the transition state structure of proteins, predicting protein stability upon mutation etc. (22,23). In this work, we have analyzed the relationship between amino acid properties and folding rates of proteins, and developed multiple regression equations for predicting protein folding rates to different structural classes of proteins using amino acid properties and a large dataset of 77 two and three-state proteins. A web server has been set up to predict the protein folding rates, which takes the amino acid sequence and structural class information as input and displays the folding rate of the protein along with amino acid composition in the output. Our method showed an excellent correlation of 0.96 between predicted and experimental folding rates of proteins. The prediction server is available online at http://psfs.cbrc.jp/fold-rate/.

MATERIALS AND METHODS

Experimental folding rates

The experimental folding rates of 77 two and three-state proteins used in related works (13,18,19) form the basis for the present study. The Protein Data Bank codes (24) and experimental $\ln(k_f)$ values are given in Table 1. The structural classification of these proteins yielded 16 all-α (dominated by α-helices; $\alpha > 40\%$ and $\beta < 5\%$), 26 all-β (dominated by β-strands; $\beta > 40\%$ and $\alpha < 5\%$) and 35 mixed class proteins (contain both α-helices and β-strands; $\alpha > 15\%$ and $\beta > 10\%$).

*To whom correspondence should be addressed. Tel: +81 3 3599 8046; Fax: +81 3 3599 8081; Email: michael-gromiha@aist.go.jp

Table 1. Predicted folding rates in a set of 77 two and three-state proteins

PDB code	Experimental[a]	ln(k_f) predicted	Deviation
All-α proteins			
1LMB	8.50	8.45	0.05
2ABD	6.55	6.33	0.22
1IMQ	7.31	7.20	0.11
2PDD	9.80	9.54	0.26
1HRC	8.76	8.66	0.10
1YCC	9.62	9.74	−0.12
256B	12.20	12.42	−0.22
1VII	11.52	11.47	0.05
1BDD	11.75	11.72	0.03
1l8W	1.61	1.61	0.00
1ENH	10.53	10.49	0.04
1EBD	9.68	9.90	−0.22
1A6N	*1.10*	*1.20*	*−0.10*
1CEI	*5.80*	*5.95*	*−0.15*
2CRO	*3.70*	*3.82*	*−0.12*
2A5E	*3.50*	*3.43*	*0.07*
All-β proteins			
1NYF	4.54	4.34	0.20
1PKS	−1.05	−0.62	−0.43
1SHG	1.41	1.57	−0.16
1SRL	4.04	4.09	−0.05
1FNF-9	−0.91	−0.96	0.05
1TEN	1.06	1.22	−0.16
1WIT	0.41	0.18	0.23
1CSP	6.98	6.75	0.23
1MJC	5.24	5.70	−0.46
2AIT	4.20	4.05	0.15
1PNJ	−1.10	−1.77	0.67
1SHF	4.50	4.78	−0.28
1C9O	7.20	7.24	−0.04
1G6P	6.30	6.11	0.19
1LOP	6.60	6.57	0.03
1PIN	9.44	9.63	−0.19
1C8C	6.91	7.04	−0.13
1PSF	*3.22*	*3.53*	*−0.31*
1FNF-10	*5.48*	*5.36*	*0.12*
1HNG	*2.89*	*2.60*	*0.29*
1HX5	*0.74*	*0.75*	*−0.01*
1TIT	*3.47*	*3.57*	*−0.10*
1IFC	*3.40*	*3.38*	*0.02*
1EAL	*1.30*	*1.20*	*0.10*
1OPA	*1.40*	*1.07*	*0.33*
1CBI	*−3.20*	*−2.68*	*−0.52*
Mixed-class proteins			
1APS	−1.48	−1.18	−0.30
1HDN	2.70	2.45	0.25
1URN	5.73	5.36	0.37
2HQI	0.18	0.47	−0.29
1PBA	6.80	6.92	−0.12
1UBQ	7.33	6.63	0.70
2PTL	4.10	3.77	0.33
1FKB	1.46	1.23	0.23
1COA	3.87	3.73	0.14
1DIV	6.58	6.87	−0.29
2VIK	6.80	6.39	0.41
1CIS	3.87	3.30	0.57
1PCA	6.80	6.68	0.12
1HZ6	4.10	4.64	−0.54
1PGB	6.00	5.75	0.25
2CI2	3.90	3.88	0.02
1AYE	6.80	7.62	−0.82
1RIS	5.90	6.08	−0.18
1POH	2.70	2.45	0.25
1BRS	3.40	3.20	0.20
1UBQ	5.90	6.63	−0.73
2ACY	0.92	1.07	−0.15
3CHY	*1.00*	*1.08*	*−0.08*
2RN2	*0.10*	*0.28*	*−0.18*
1RA9	*−2.50*	*−2.30*	*−0.20*
1BNI	*2.60*	*2.90*	*−0.30*
2LZM	*4.10*	*4.14*	*−0.04*
1SCE	*4.20*	*4.46*	*−0.26*
1GXT	*4.38*	*4.15*	*0.23*
2A5E	*3.50*	*3.71*	*−0.21*
1AON	*0.80*	*1.46*	*−0.66*
1PHP (N-TERMINAL)	*2.30*	*2.02*	*0.28*
1PHP (C-TERMINAL)	*−3.45*	*−3.71*	*0.26*
1qop (α-Subunit)	*−2.53*	*−2.76*	*0.23*
1qop (β-Subunit)	*−6.91*	*−6.86*	*−0.05*

[a]Experimental folding rates are obtained from Galzitskaya *et al*. (13), Ivankov and Finkelstein (18) and Gromiha (19). The three-state proteins are shown in italics.

Amino acid properties

We used a set of 49 diverse amino acid properties (physical–chemical, energetic and conformational), which fall into various clusters analyzed by Tomii and Kanehisa (25) in the present study. The amino acid properties were normalized between 0 and 1 using the expression, $P_{norm}(i) = [P(i) - P_{min}]/[P_{max} - P_{min}]$, where $P(i)$, $P_{norm}(i)$ are, respectively, the original and normalized values of amino acid i for a particular property, and P_{min} and P_{max} are, respectively, the minimum and maximum values. The numerical and normalized values for all the 49 properties used in this study along with their brief descriptions have been explained in our earlier articles (26,27).

Computational procedure

The average amino acid property for each protein, $P_{ave}(i)$ was computed using the standard formula,

$$P_{ave}(i) = \sum_{j=1}^{N} P(j)/N \qquad \mathbf{1}$$

where, $P(j)$ is the property value of jth residue and the summation is over N, the total number of residues in a protein. The computed property value $P_{ave}(i)$ for each class of proteins was related with experimental folding rate $\ln k_f(i)$ using single correlation coefficient. Further, we have combined the amino acid properties using multiple regression technique (28). It is the extension of linear single regression with more than one property. For example, the multiple regression equation for $\ln k_f$ with two properties, P_1 and P_2 may be written as: $\ln k_f = c + aP_1 + bP_2$, where c is a constant, and a and b are regression coefficients obtained by fitting the properties P_1 and P_2 with $\ln k_f$.

RESULTS AND DISCUSSIONS

Prediction of protein folding rates

We have analyzed the relationship between amino acid properties and folding rates of proteins belonging to different structural classes. We observed that the conformational and thermodynamic properties show good correlation with protein folding rates in all-α proteins. In all-β proteins, thermodynamic properties and in mixed class proteins,

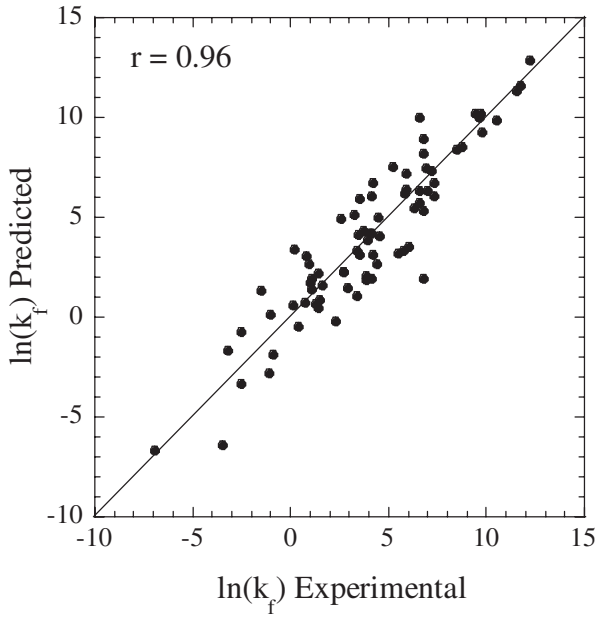

Figure 1. Relationship between experimental and predicted ln(k_f) values using multiple regression model with jack-knife test in a set of 77 two and three-state proteins.

physical–chemical properties show significant correlation with protein folding rates. This observation is similar to that reported earlier with a small set of proteins (19).

We have combined the amino acid properties using multiple regression technique. The combination of 10 properties in all-α proteins yielded the correlation coefficient of 0.997 and the predicted folding rates are presented in Table 1. We have also performed jack-knife test to examine the validity of the present method and the results are shown in Figure 1. In this test, the folding rates of ($n - 1$) proteins have been used to derive the coefficients in multiple regression equation and the same coefficients have been used for predicting the folding rate of the left-out protein. The same procedure has been repeated for n times for obtaining the folding rates of all proteins. We observed an excellent correlation of 0.992 between experimental and predicted folding rates of proteins in the considered all-α proteins. We have also carried out the computations for all-β and mixed class proteins and the results are included in Table 1 and Figure 1. We obtained the correlation of 0.99 for both all-β and mixed class proteins in back-check prediction. The jack-knife test yielded the correlation of 0.97 and 0.90, respectively for all-β and mixed class proteins. The less correlation of mixed class proteins than all-α and all-β classes of proteins might be due to the complicate nature of the structures with both α-helices and β-strands.

Considering all the 77 proteins together, we obtained the overall correlation of 0.99 and 0.96, respectively, for the back-check prediction and jack-knife tests. The deviation between experimental and predicted folding rates is 0.23 for the back-check prediction and 1.19 for the jack-knife test. Further, the prediction results obtained with two-state proteins are similar to that obtained with three-state proteins. It may be noted that the reported folding rates are given in logarithmic scale and the actual error will be relatively high in the real values of protein folding rates. Interestingly, the folding rates of 44 out of 77 proteins (57%) are predicted within the deviation of less than 1 unit. In the self-consistency test (back-check prediction) all proteins have been predicted within this limit. Further, we have developed a single equation for predicting the folding rates of proteins without structural class information and obtained the correlation of 0.87.

Prediction on the web

We have developed a web server for predicting the folding rates of two and three-state proteins. Figure 2a shows the details of our web server including the input options. It takes the amino acid sequence in one letter format as the input and automatically omits gaps. It also gets the information about the structural class. The structural class information for a protein of known structure can be obtained either from SCOP or CATH databases and prediction results for structural class can be obtained with other servers, such as, protein structure prediction server (PSA; http://bmerc-www.bu.edu/psa/), secondary structural content prediction (SSCP; http://www.bork.embl-heidelberg.de/SSCP/) etc. The output formats are shown in Figure 2b. It shows the amino acid composition of the query protein, selected type of the protein and the predicted folding rate. As an example, for λ repressor belonging to all-α protein, the predicted folding rate, ln(k_f), is 8.44/s, which agrees remarkably well with experimental observations (8.50/s). The prediction results are freely available at http://psfs.cbrc.jp/fold-rate/.

Comparison with other methods

The methods based on 3D structures of proteins reveals the relationship between structural parameters, such as contact order, long-range order, etc. and protein folding rates. These methods showed a correlation in the range of 0.8–0.9. Gong *et al.* (17) reported the correlation of 0.91 using secondary structure content. Recently Putna and Rost (20) predicted protein folding rates from amino acid sequence using the information about predicted long-range contacts and reported a correlation of 0.61 for a set of 37 proteins. The present method shows the correlation of 0.97 and 0.93 between experimental and predicted folding rates with the back-check and jack-knife tests, respectively. These accuracy levels are better than other methods in the literature. Although the direct comparison of correlation coefficients obtained in the present work with the other methods is not appropriate, the empirical relationships derived for different structural classes predict the folding rates with high accuracy.

Limitations of the present method and possible improvements

In the present work, we have used the amino acid sequence and 49 amino acid properties for predicting protein folding rates using multiple regression equations. We observed a good agreement between predicted and experimental folding rates of proteins. It may be noted that the folding rates of proteins depend on experimental conditions, which are not considered in the present work. On the other hand, the inclusion of structural and evolutionary information may improve the prediction accuracy.

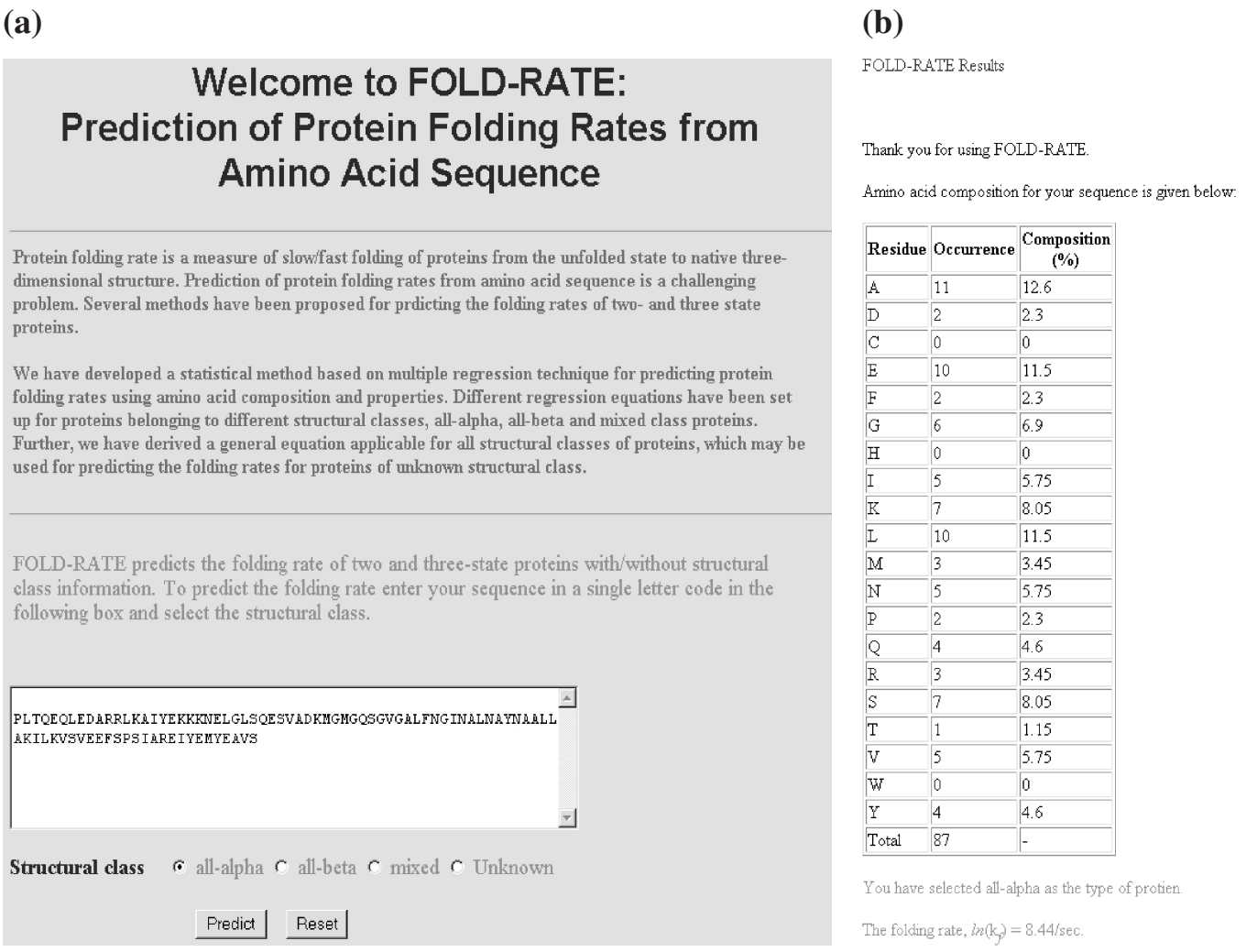

Figure 2. Web based prediction of protein folding rates. (**a**) First page showing the input format (amino acid sequence in single letter code; an example is shown for λ repressor (1LMB) and structural class information (all-α). (**b**) The query sequence, amino acid composition, type of the protein and predicted folding rate are shown. The ln(k_f) value is predicted to be 8.44/s.

CONCLUSIONS

We have devised a method based on multiple regression technique for predicting protein folding rates using amino acid properties. Different regression equations have been developed for proteins belonging to all-α, all-β and mixed class proteins. We observed that the predicted folding rates show an excellent agreement with experimental results. A web server has been developed for the prediction purpose and the results are available online, which may be very helpful for the users to get the folding rate of any protein with its structural class information.

ACKNOWLEDGEMENTS

The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Eaton,W.A., Munoz,V., Hagen,S.J., Jas,G.S., Lapidus,L.J., Henry,E.R. and Hofrichter,J. (2000) Fast kinetics and mechanisms in protein folding. *Ann. Rev. Biophys. Biomol. Struct.*, **29**, 327–359.
2. Plaxco,K.W., Simons,K.T. and Baker,D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.
3. Debe,D.A. and Goddard,W.A.,III (1999) First principles prediction of protein folding rates. *J. Mol. Biol.*, **294**, 619–625.
4. Gromiha,M.M. and Selvaraj,S. (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.*, **310**, 27–32.
5. Munoz,V. and Eaton,W.A. (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA*, **96**, 11311–11316.
6. Dinner,A.R. and Karplus,M. (2001) The roles of stability and contact order in determining protein folding rates. *Nature Struct. Biol.*, **8**, 21–22.
7. Makarov,D.E., Keller,C.A., Plaxco,K.W. and Metiu,H. (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl Acad. Sci. USA*, **99**, 3535–3539.
8. Zhou,H. and Zhou,Y. (2002) Folding rate prediction using total contact distance. *Biophys. J.*, **82**, 458–463.

9. Makarov,D.E. and Plaxco,K.W. (2003) The topomer search model: a simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.*, **12**, 17–26.
10. Dokholyan,N.V., Li,L., Ding,F. and Shakhnovich,E.I. (2002) Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, **99**, 8637–8641.
11. Zhang,L., Li,J., Jiang,Z. and Zia,A. (2003) Folding rate prediction on neural network model. *Polymer*, **44**, 1751–1756.
12. Gromiha,M.M. (2003) Importance of native state topology for determining the folding rate of two-state proteins. *J. Chem. Inf. Comp. Sci.*, **43**, 1481–1485.
13. Galzitskaya,O.V., Garbuzynskiy,S.O., Ivankov,D.N. and Finkelstein,A.V. (2003) Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins*, **51**, 162–166.
14. Ivankov,D.N., Garbuzynskiy,S.O., Alm,E., Plaxco,K.W., Baker,D. and Finkelstein,A.V. (2003) Contact order revisited: influence of protein size on the folding rate. *Protein Sci.*, **12**, 2057–2062.
15. Shao,H., Peng,Y. and Zeng,Z.H. (2003) A simple parameter relating sequences with folding rates of small alpha helical proteins. *Protein Pept. Lett.*, **10**, 277–280.
16. Micheletti,C. (2003) Prediction of folding rates and transition-state placement from native-state geometry. *Proteins*, **51**, 74–84.
17. Gong,H., Isom,D.G., Srinivasan,R. and Rose,G.D. (2003) Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.*, **327**, 1149–1154.
18. Ivankov,D.N. and Finkelstein,A.V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 8942–8944.
19. Gromiha,M.M. (2005) A statistical model for predicting protein folding rates from amino acid sequence with structural class information. *J. Chem. Inf. Model*, **45**, 494–501.
20. Punta,M. and Rost,B. (2005) Protein folding rates estimated from contact predictions. *J. Mol. Biol.*, **348**, 507–512.
21. Gromiha,M.M. and Selvaraj,S. (2004) Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.*, **86**, 235–277.
22. Gromiha,M.M., Oobatake,M., Kono,H., Uedaira,H. and Sarai,A. (1999) Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng.*, **12**, 549–555.
23. Gromiha,M.M. and Selvaraj,S. (2002) Important amino acid properties for determining the transition state structures of two-state protein mutants. *FEBS Lett.*, **526**, 129–134.
24. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
25. Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.
26. Gromiha,M.M., Oobatake,M. and Sarai,A. (1999) Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys. Chem.*, **82**, 51–67.
27. Gromiha,M.M., Oobatake,M., Kono,H., Uedaira,H. and Sarai,A. (2000) Importance of surrounding residues for protein stability of partially buried mutations. *J. Biomol. Str. Dyn.*, **18**, 281–295.
28. Grewal,P.S. (1987) *Numerical Methods of Statistical Analysis*. Sterling publishers, New Delhi.